Computational Challenges in Data Citation

University of Pennsylvania, April 17-18, 2014

Workshop Report

Peter Buneman, Sarah Cohen-Boulakia, Susan B. Davidson, Jim Frew, Val Tannen

Introduction

Citation is an essential part of scientific publishing and, more generally, of scholarship. It is used to gauge the trust placed in published information and, for better or for worse, is an important factor in judging academic reputation. Now that so much scientific publishing involves data and takes place through a database rather than conventional journals, how is some part of a database to be cited? More generally, when one extracts some data from a large, complex, evolving database, how does one create the appropriate citation? How does one verify that the citation is correct?

Frameworks have been put forward by Information Scientists to serve as models or templates for citation. At the same time Data Scientists associated with various disciplines such as Bioinformatics, Earth Sciences, Neuroscience, etc., encounter interesting problems in trying to foster the citation of data. However, it is clear that for large evolving datasets and databases we are going to need algorithmic techniques and software technologies both to generate and to verify the correctness of citations, and these may well pose new problems for Computer Scientists.

The purpose of this workshop was to bring together people representing these different disciplines and enumerate the computational challenges and opportunities associated with data citation. The workshop was organized around three sessions – Citation Principles and Standards, Citation and Linked Open Data, and Executable Papers and Reproducibility – during which an overview talk was given followed by perspectives by participants. Participants then broke out into breakout groups, each of which contained people from different disciplines, and brainstormed what they believed to be the most important computational challenges for data citation. During a plenary session the next day, the challenges were revisited and refined. This report represents these findings.

In the remainder of this report, we discuss what data citations are and how they differ from citations to printed material as well as links. We then present the key computational challenges

which emerged from discussions at the workshop. The appendix contains information about workshop participants, as well as references to material on topics related to the workshop.

What is a data citation and why is it special?

Citations are the fabric of the "scholarly record" or the "scientific canon". They are used to retrieve the cited material, give credit to the creator of the material, assign responsibility or ownership to the material, and enable one to know who is to be consulted on issues of intellectual property or privacy.

Two central and related concepts behind citation are *permanence* and *authentication*. Printed publications, by their nature, are static and are "guaranteed" to exist in the future as originally published by their publisher; new versions are created relatively slowly and are managed by version numbers. In contrast, data in a digital environment may have no guarantor whose reputation is tied to the data existing at a future point in time; changes to the data may be relatively small and occur very frequently; and the schema of the database may evolve in addition to its contents. Guaranteeing that cited data will exist in the future, that it has not been tampered with (authentication), and that it can be reached by the given citation are challenging problems for digital data.

The notion of *hierarchy* of citations is already present in a print environment, examples of which include a chapter within a book, or a paper within the proceedings of a yearly conference. However, the hierarchies are relatively shallow and there are well-understood templates for citations. With digital data, the hierarchies may be much deeper and the conventions for citation undefined. For example, the data may be a hierarchical database (a tree). While the database as a whole may be credited to some set of individuals (the owners), many different people may have contributed content at different levels in the tree. The root of the database may therefore be a top-level citation, but subtrees of the database may also be citable. From any given node in the tree there may therefore be a chain of citations up to the root. The owner of the database may therefore wish to be able to specify citation conventions for various portions of the tree to credit and incentivize contributors. This issue becomes even more complex when the data is graph-structured, e.g. RDF data.

Related to the hierarchical structure of data citations is the notion of a *citable unit.* While the cited data may be low down in the hierarchy, the node that contains the relevant information including metadata and context may be higher up. This is the citable unit, and a citation may carry both the location of the unit and the location of the data within that unit. In other situations, there may be no hierarchy, or the metadata and context may exist in a separate place and require a different access method.

Traditional citations not only carry some means of locating the cited material such as a journal issue, a digital object identifier or a URI; they also carry extra “snippets” of information such as title, authorship, date etc. These snippets serve various functions; in particular, they are important for (human) recognition of the cited material. What these snippets are, what functions they perform, and how they differ from location information are questions that may have different answers that depend on what is being cited. However, it is clear that in addition to data citations being *machine actionable* they should allow *human recognition*.

An important distinction from the traditional setting is that data being cited may be obtained by some *query* over or *view* of the database. The result may therefore be a set of data items, each of which could have an independent citation. In addition to the potentially *large number* of the resulting data citations, data in the result are related by being in the answer set of the same query. Therefore, the query itself may be an important component of the citation and should be impervious to schema restructuring. Another consequence of query-oriented access to cited data is that the same data set may be reachable from (generated by) several different queries, raising the question of what the appropriate citation for the data set should be and whether it should be unique.

New Roles for Citation

We tend to think, rather conservatively, of a citation as being something one puts in a paper. However we need to take into account that the reasons for using citations have changed rather rapidly over the past two decades, and that the nature of scientific papers is also changing. It is difficult to give a coherent summary of this evolution, but we list some of the issues here.

- **Workflows, executable papers and provenance.** A huge number of scientific papers include the results of computational analysis pipelines. This has led to the notion of an *executable paper* which somehow contains enough information for the analysis to be repeatable. This is also obviously related to provenance-enabled scientific workflows which are designed, in part, to ensure repeatability. Provenance information may also involve links to data. An interesting question is whether these links should be associated with conventional citation?

- **Citations and Linked Open Data.** There is an interesting relationship between citations and Linked Open Data (LOD). LOD embodies a set of best practices for “exposing, sharing, and connecting data, information and knowledge on the Semantic Web using URIs and RDF” (Wikipedia). Persistent links (URIs) are an important part of citation, but we need more as described earlier -- information such as authorship, title, context. We may also want to be able to cite LOD, and give credit to the people who created it. Unlike many of the earlier examples, which have a well-organized hierarchical structure, RDF is graph structured.

- **Citation sociology**. It is only rather recently that citation counts have been used as a measurement of academic merit, but it not clear that a straightforward citation count is the right estimate. Should we be using something more like PageRank[1]? Should we take into account the reason that a citation is there, perhaps through some form of citation ontology[2]? There is an argument that the number of intellectual influences on a paper is necessarily small, and that therefore we should have a two-tier system in which influential citations appear "in-line" in a paper, while others are confined to a "methods" or "related work" appendix.

- **Citation content**. Although there are a huge number of formats for citations, the content of a citation is relatively fixed, and standards for data citation have been created by "tweaking" existing conventions. The form and content of the citation is up to the person doing the citing, but it is relatively straightforward. We are now seeing, especially in curated databases, the producers of the data specifying the citation.

None of these issues diminishes the importance of citations, and some introduce even more complexity into the computational issues surrounding data citations

Computational Challenges

A number of computational challenges arise from the special character of data citations, which we now enumerate. In what follows, we will refer to different categories of users of citations: *exporters* (the people who manage/produce the data), *citers* (the people who cite the data), and *consumers* (the people who access the data being cited).

1. **Modeling the referent of a citation**: There are a large number of standards for traditional citations, each of which is structured around well-accepted concepts (e.g. booktitle, title, author, page, year, …). However, data citations do not fit into this structure, and exporters must have a formal framework for specifying what a citation to data in their database should look like. The framework should be general enough to be used by different fields, and to capture many different types of referents (set of tuples, bit stream, subgraph, etc). The framework should be understandable by exporters, citers, as well as consumers. Subproblems include:
    a. **Level of granularity**: if data is hierarchically structured (e.g. attribute, tuple, set of tuples) what level should the data be cited at? Can the level be inferred from the structure?

[1] en.wikipedia.org/wiki/**PageRank**

[2] David Shotton. CiTO, the Citation Typing Ontology. Bio-Ontologies 2009: Knowledge in Biology Stockholm, Sweden. 28 June 2009

b. **Syntactic/physical representation**: the same data can be represented or serialized in different ways, and the representation has an effect on the outcomes of future processing (e.g. the representation may be lossy). If so, the representation of the data must be captured in the model.

2. **Automatic generation of citations**: Since databases are large, data citations should be automatically generated using the database's citation specification. This is of particular importance to exporters, but as shown in the subproblems below is also essential to citers and consumers.
    a. **Citing data and process**: The process by which the data being cited is arrived at may be important, e.g. if it is generated by a query to the database. Should the citation give the data (extensional) or a way of obtaining the data (intensional)?
    b. **Creating views of large citation sets**: Citations may be generated and collected as part of the execution of an ensemble of workflows, and could therefore become extremely numerous. What are ways of refining or abstracting extremely large sets of citations? For example, are some citations more important than others in the context of the process by which they were generated, or is it possible to generalize in the citation hierarchy? The original citations may still exist, but a "view" of the large set could be created with abstractions of zooming in/out to the desired level of detail.

3. **Verification of citations**: Given a citation C and some referent M, how do we verify that M is what was cited by C? Or given just C, can we detect if its referent can exist? Related issues:
    a. **Detection of referent relationship**: If it is possible for data to have multiple citation representations, can we detect when two different citations refer to the same data? Or can we detect when one referent is a component of the other? Note that this is an important problem for refining large sets of citations.
    b. **Closures of citations**: Most data has been created as a process involving other data as inputs. One can therefore talk about the closure of a data citation as being citations to the data from which it has been produced (analogous to the closure of the references of a paper). Given a citation C and some data M, one may therefore want to know if M is cited transitively by the referent of

4. **Making data citable**: Many databases already exist and are citable, for example, by reference to a traditional publication. However, suppose the owners now wish to create citable units within the database? In addition to specifying how citations are to be generated and doing so automatically (points 1 and 2 above), additional issues to be addressed include:
    a. **Fixity**: The referent must be recoverable from the database despite changes in content and schema. One approach would be to archive different snapshots of the data, but when is this done (what are the triggers) and can it be done in a scalable and sustainable way?

b. **Chain of trust**: What can be done to guarantee that the database meets the requirements of being citable? This is related to verification of citations, but there may also be an idea of a digital signature that can be compared against the referent to see if it is the same with high probability.

5. **Grouping and Structuring Citations**: Within a paper, citations serve different purposes and may need to be arranged to suit these purposes. For example, the major intellectual influences on a paper may be relatively few, and are often cited in-line, while there may be a huge number of citations needed to support some computational analysis. The former are often structured for human readability while the latter may be structured for machine readability and reproducibility. In the latter case, to what extent can the arrangement be derived from, say, the workflows or other computational systems that were used to derive the results?

Conclusions

Although numerous committees have done a great deal of work on advocating data citation and establishing principles and standards, there is the social challenge of getting the practice of data citation commonly accepted and the computational challenge of making it easy to do. The challenges are pervasive. We need tools for those who produce and publish data so that they can make their data citable and say how they want it to be cited; we need tools for those who preserve or archive data so that citations remain effective and verifiable; and we need tools for those who write papers in order that citation becomes an integral part of the process of making their results reproducible. Most importantly, users who wants to cite some data should not be given principles and standards, but an appropriate citation that conforms to those standards.

Appendix A: Workshop Participants

Micah Altman	MIT Library, Director of Research
Bruce Barkstrom	Former NASA Data Center Manager
Peter Buneman	Professor of Database Systems, University of Edinburgh
Wang Chiew Tan	Professor of Computer Science, UC Santa Cruz
Fernando Chirigati	Ph.D. candidate, NYU Polytechnic School of Engineering
Sarah Cohen-Boulakia	Assistant Professor, LRI Bioinformatics Group, Paris-Sud University
Susan Davidson	Professor of Computer and Information Science, University of Pennsylvania
Ruth Duerr	Data Scientist/Systems Engineer, NSIDC
Wenfei Fan	Professor of Database Systems, University of Edinburgh
Juliana Freire	Professor of Computer Science and Engineering, Center for Data Science, NYU
James Frew	Associate Professor, Bren School of Environmental Science and Management, UCSB
Devarshi Ghoshal	Ph.D. candidate, School of Informatics and Computing, Indiana University
Carole Goble	Professor of Computer Science; co-Director, Information Management Group, University of Manchester
Paul Groth	Assistant Professor, Web and Media Group; VU Amsterdam
Margaret Hedstrom	Professor, University of Michigan, School of Information
Bill Howe	Associate Director, e-Science Institute, University of Washington
Zack Ives	Associate Professor of Computer and Information Science, University of Pennsylvania
HV Jagadish	Professor of Electrical Engineering and Computer Science, University of Michigan
Greg Janée	Specialist, Earth Science Institute, UCSB
Bertram Ludaescher	Professor of Computer Science, Department of Computer Science and the Genome Center, UC Davis
Maryann Martone	Professor-in-Residence, National Center for Microscopy and Imaging, UCSD Health Sciences
John Ockerbloom	Digital Library Architect, University of Pennsylvania

Seamus Ross	Dean, Faculty of Information, University of Toronto.
Joanna Sharman	Database Development Manager, IUPHAR-DB
Val Tannen	Professor, University of Pennsylvania
Curt Tilmes	Computer Scientist, NASA Goddard Space Flight Center
Herbert Van de Sompel	Lead, Digital Library Research and Prototyping Team, LANL

Appendix B: Reference Material

Standards and Principles

- FORCE11 Joint Declaration on Data Citation Principles
- Out of Cite, Out of Mind: The Current State of Practice, Policy, and Technology for the Citation of Data
- Scalable Dynamic Data Citation (RDA WG-DC Position Paper)
    - RDA Working Group on Data Citation (WG-DC)
- For Attribution—Developing Data Attribution and Citation Practices and Standards
- Supporting Science through the Interoperability of Data and Articles
- DOIs unambiguously and persistently identify published, trustworthy, citable online scholarly literature. Right?
- Towards a data publishing framework for primary biodiversity data: challenges and potentials for the biodiversity informatics community
- The Anatomy of a Data Citation
- Publishing: Credit where credit is due

Citation systems design and implementation

- Peter Buneman. How to Cite Curated Databases and how to Make them Citable
- Micah Altman and Mercè Crosas. The Evolution of Data Citation: From Principles to Implementation

Data Citation and Linked Data

- Linked Data: Design Issues
- Cool URIs Don't Change
    - Cool URIs and the Semantic Web
- Henry S. Thompson. Naming on the Web: What scholars should want, and what they can have
- Linked Data: Effort Distribution between Publishers, Consumers and Third Parties
- Web-scale Data Integration: You can only afford to Pay As You Go
- Versioned URLs and Memento
- A proposal for the inclusion of temporal information as attributes on links

Data Citation and Reproducibility

- Roger Peng. Reproducible Research in Computational Science
- Jennifer Schopf. Treating Data Like Software: A Case for Production Quality Data
- Accessible Reproducible Research, Jill Mesirov, Science 22 January 2010: Vol. 327 no. 5964 pp. 415-416 DOI: 10.1126/science.1179653
- Enhancing and abstracting scientific workflow provenance for data publishing
- Reproducible Science
- Reproducible Research
- Carole Goble. Results may Vary

- Results may vary: reproducibility, software, science (more compact keynote with a focus on software)
- Recomputation.org
- Code as a Research Object
- Research Objects
- A model project for reproducible papers: critical temperature for the Ising model on a square lattice.
- Packing experiments for sharing and publication.
- Computational reproducibility: state-of-the-art, challenges, and database research opportunities.
- Heather Christenson. HathiTrust: A Research Library at Web Scale
- Research Data Alliance Working Group Data Citation (WG-DC). See especially their Case Statement

Use Cases

- The IUPHAR/BPS Guide to Pharmacology
- The EBI Experimental Factor Ontology
- Recommended practices for citation of data published through the GBIF network